DS105

Machine Learning Project (Titanic)

Presented by: Darius

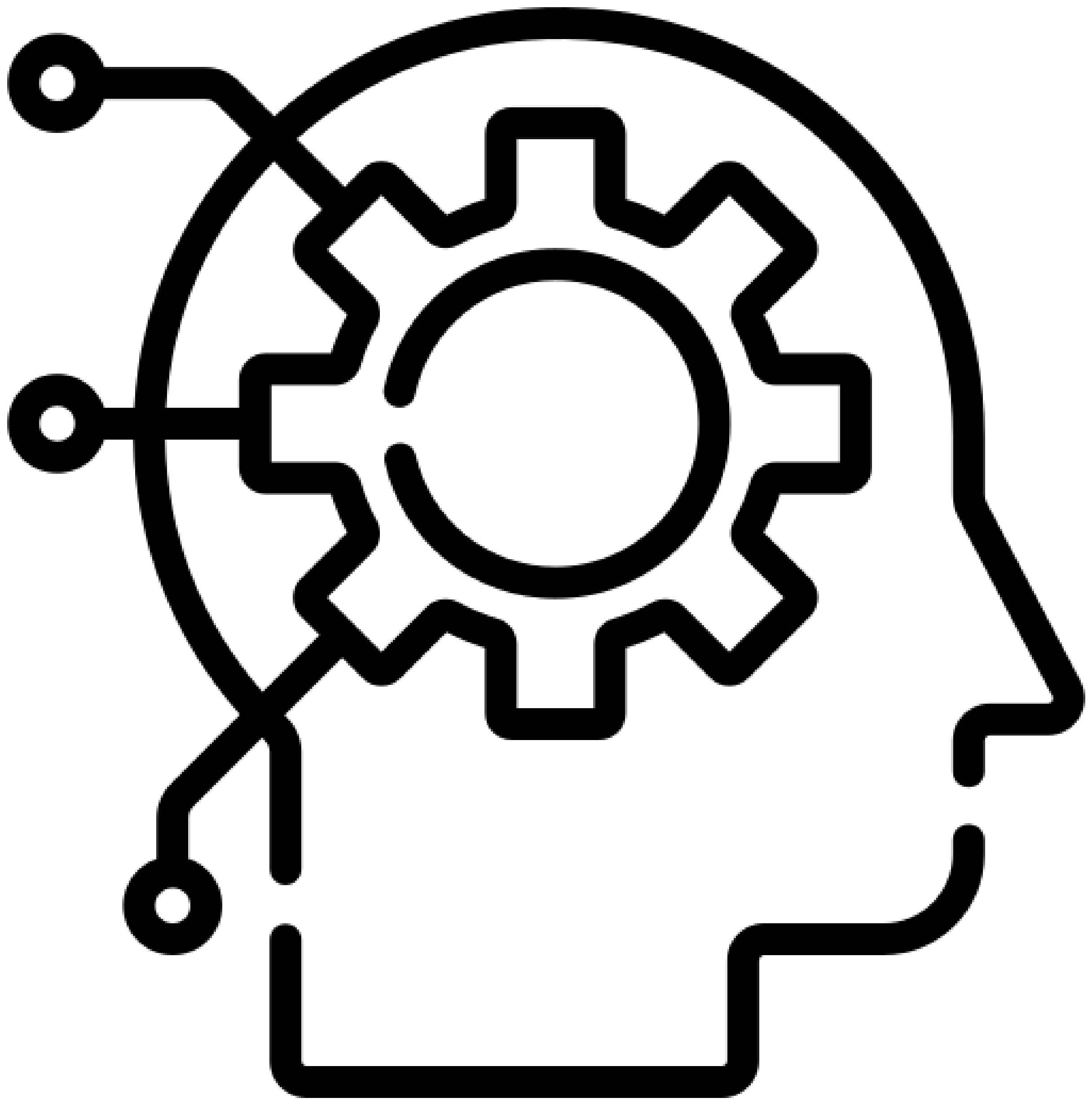
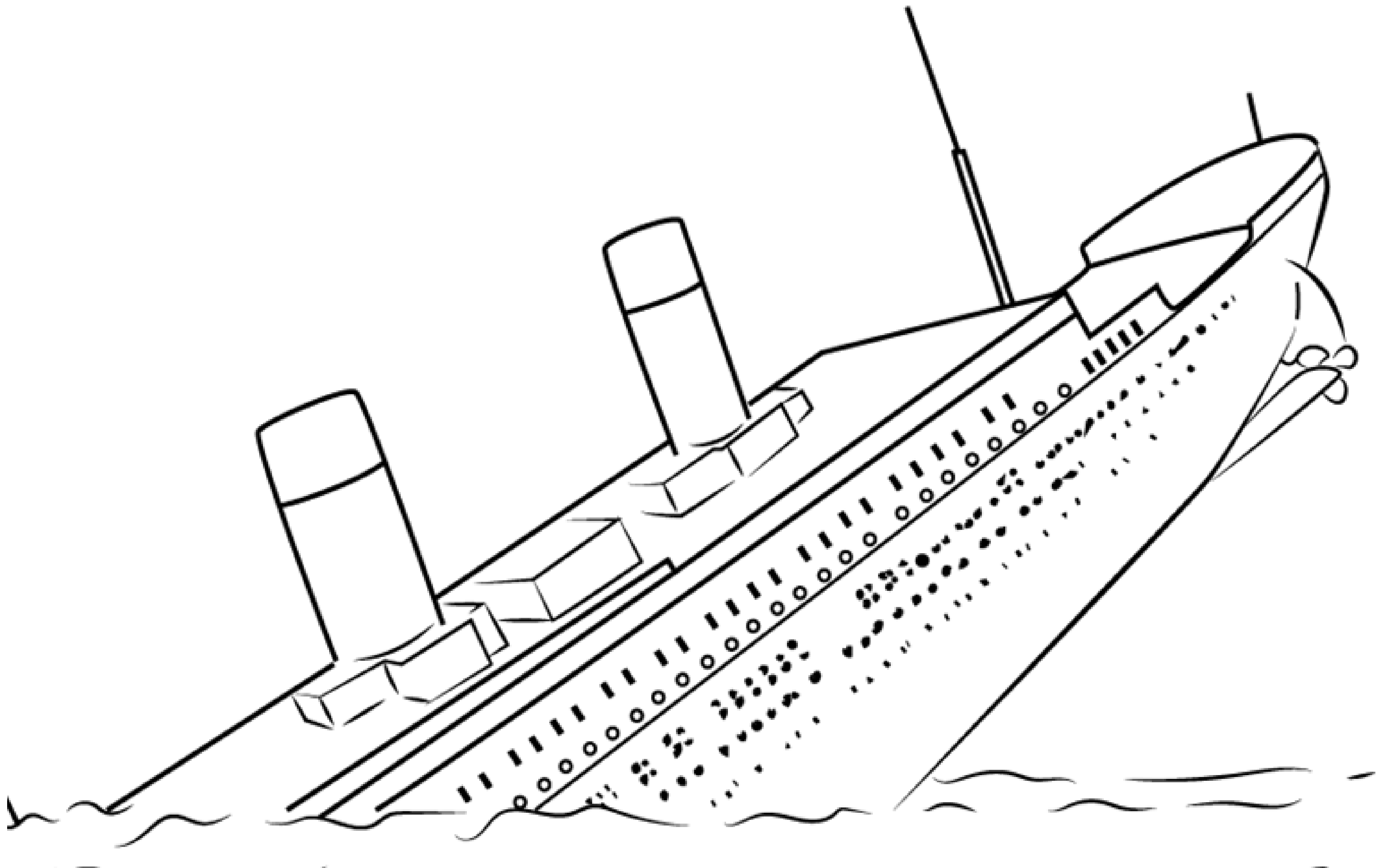
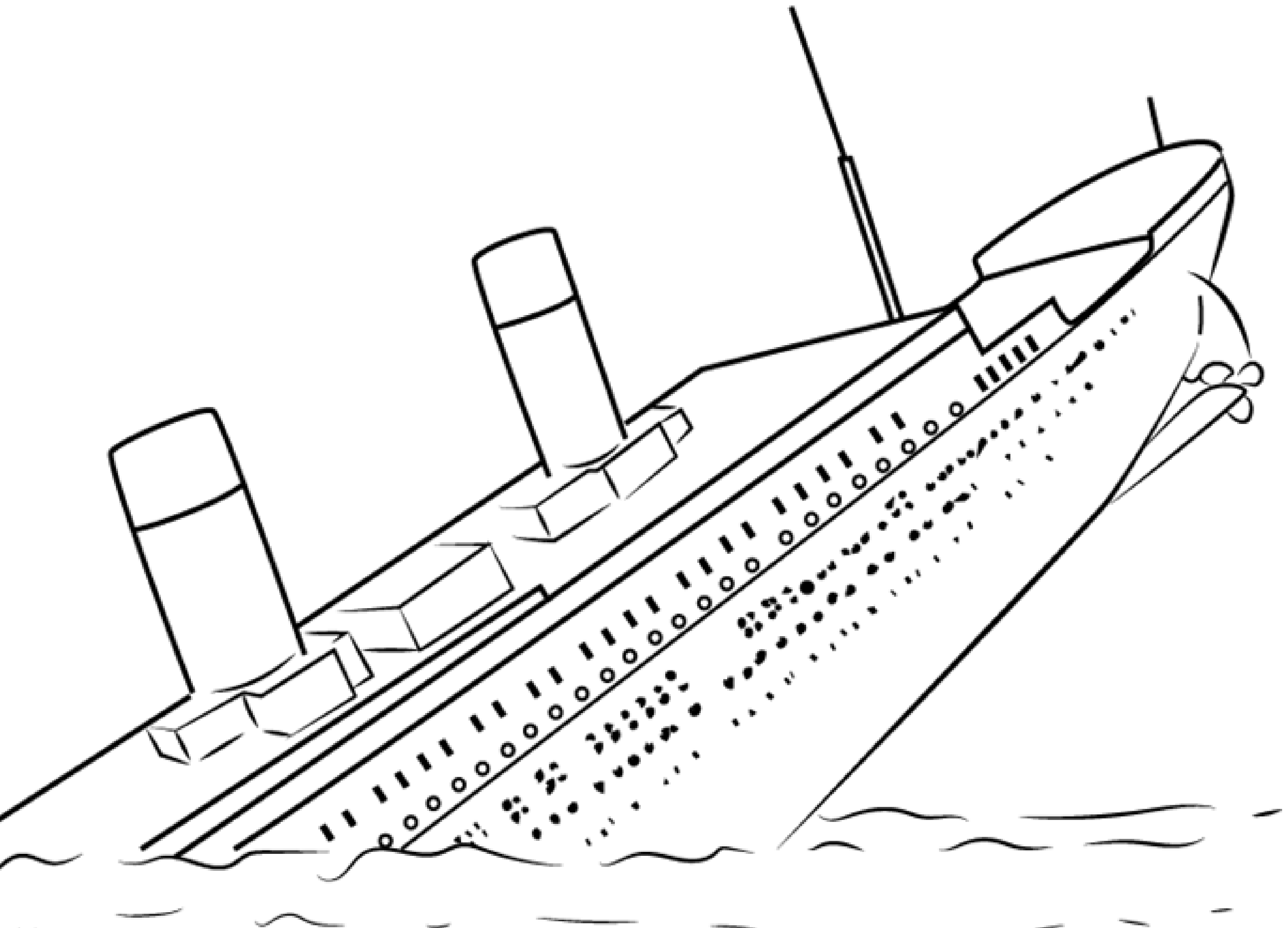


Table of Content



Recap



Problem Statement

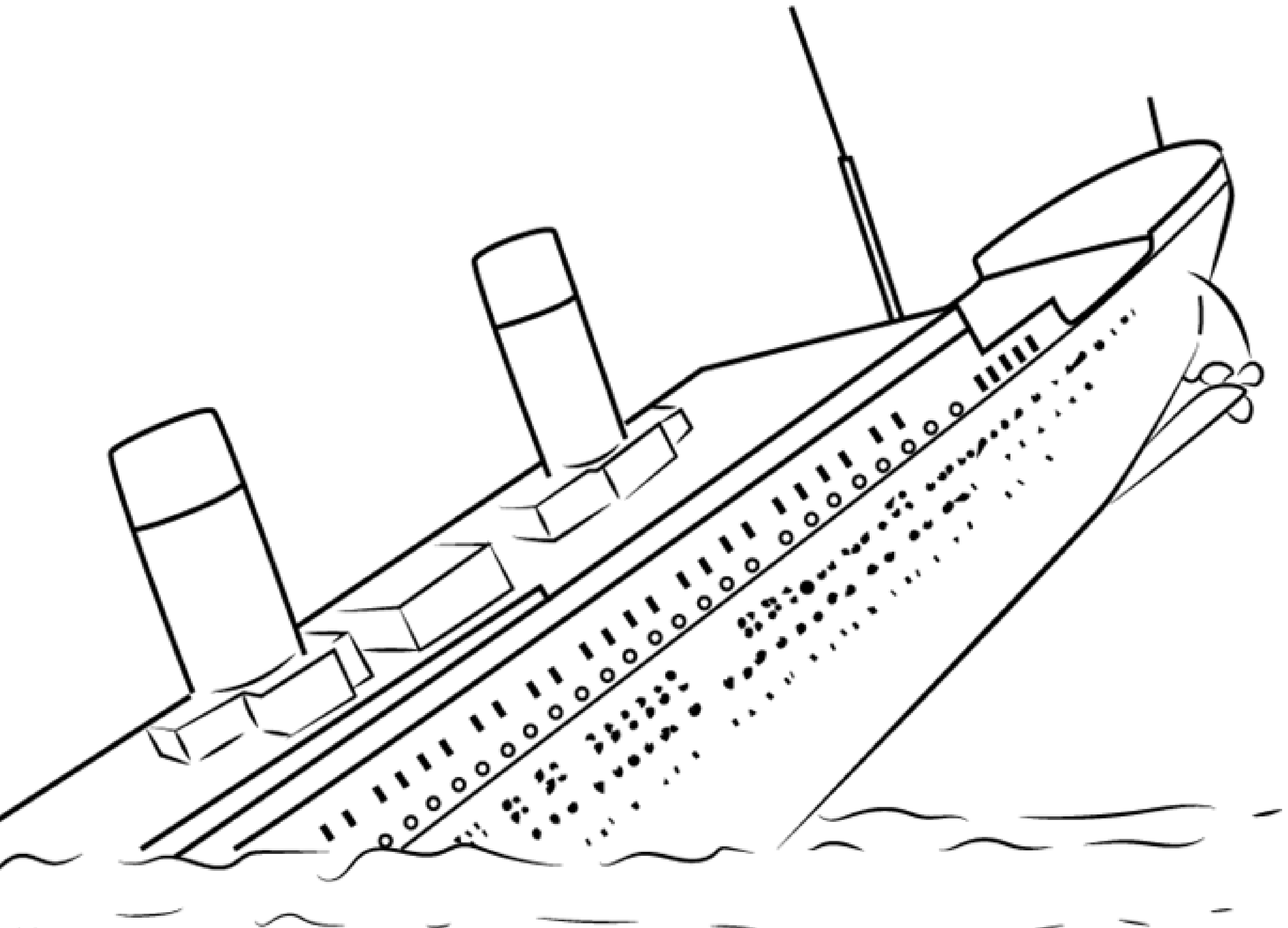
- To use different machine learning models and predict if the passengers survived the Titanic shipwreck.
- Which model is the best in this scenario

Dataset

- Total of 891 records
- 12 columns
- No duplicate data found.
- 3 datatypes in the record (int,float,object)

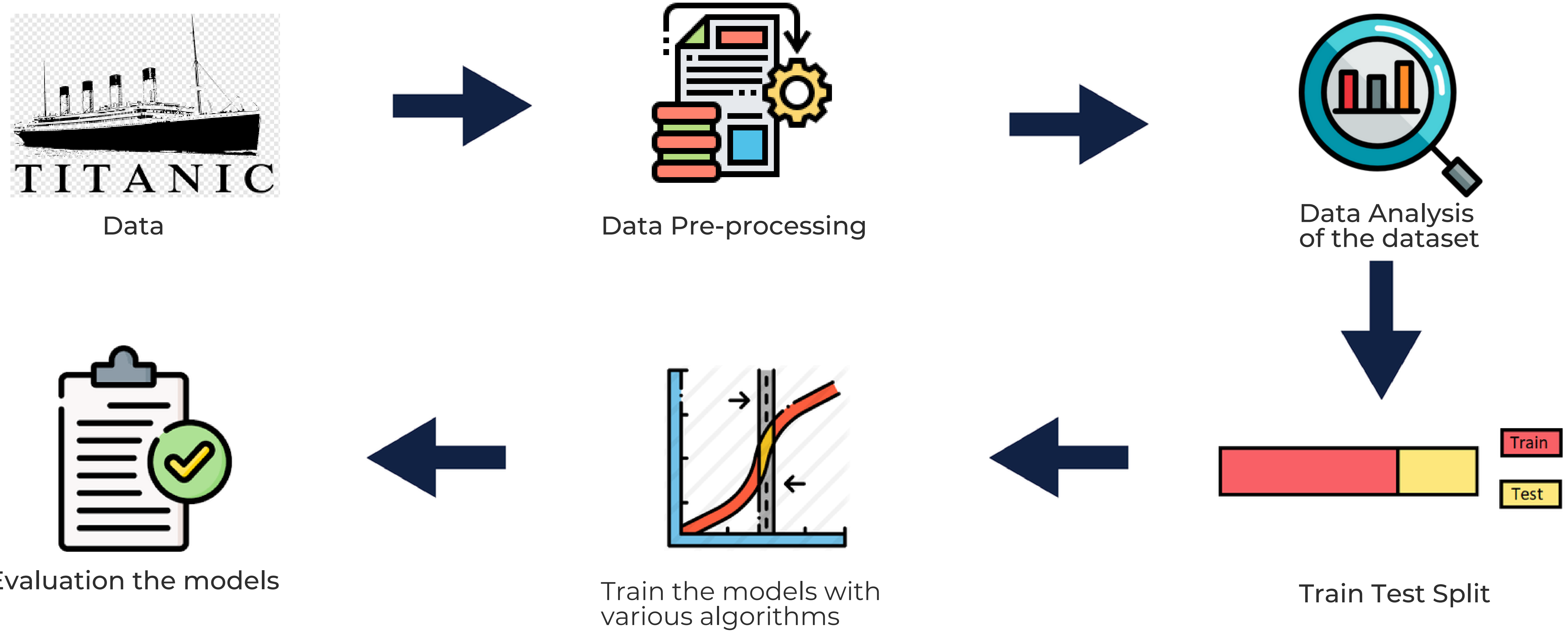
Recap

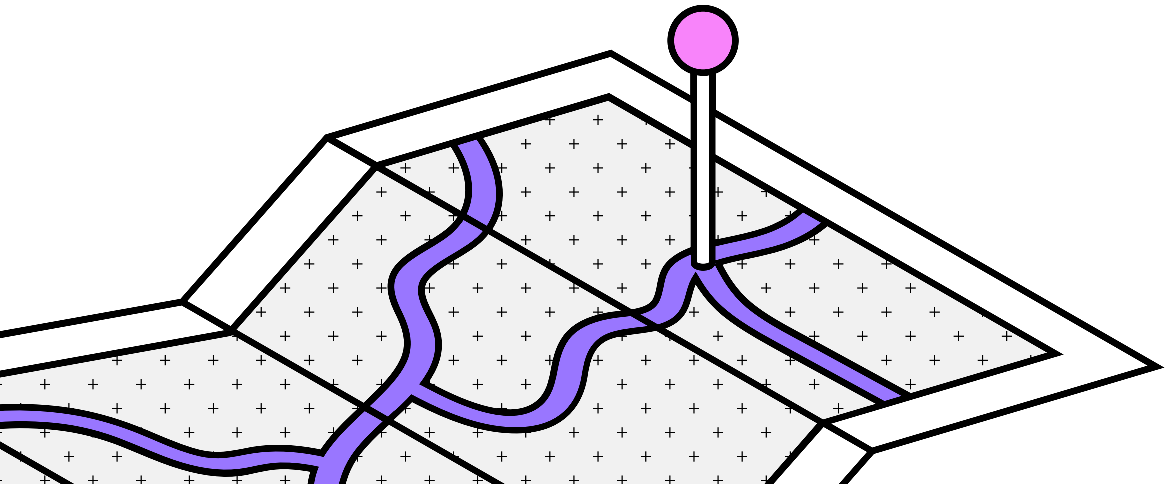
	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
5	897	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S
6	898	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q
7	899	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S
8	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C
9	901	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S
10	902	3	Ilieff, Mr. Ylio	male	NaN	0	0	349220	7.8958	NaN	S
11	903	1	Jones, Mr. Charles Cresson	male	46.0	0	0	694	26.0000	NaN	S
12	904	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23.0	1	0	21228	82.2667	B45	S
13	905	2	Howard, Mr. Benjamin	male	63.0	1	0	24065	26.0000	NaN	S
14	906	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance...	female	47.0	1	0	W.E.P. 5734	61.1750	E31	S
15	907	2	del Carlo, Mrs. Sebastiano (Argenia Genovesi)	female	24.0	1	0	SC/PARIS 2167	27.7208	NaN	C
16	908	2	Keane, Mr. Daniel	male	35.0	0	0	233734	12.3500	NaN	Q
17	909	3	Assaf, Mr. Gerios	male	21.0	0	0	2692	7.2250	NaN	C
18	910	3	Ilmakangas, Miss. Ida Livija	female	27.0	1	0	STON/O2. 3101270	7.9250	NaN	S
19	911	3	Assaf Khalil, Mrs. Mariana (Miriam")"	female	45.0	0	0	2696	7.2250	NaN	C
20	912	1	Rothschild, Mr. Martin	male	55.0	1	0	PC 17603	59.4000	NaN	C
21	913	3	Olsen, Master. Artur Karl	male	9.0	0	1	C 17368	3.1708	NaN	S
22	914	1	Flegenheim, Mrs. Alfred (Antoinette)	female	NaN	0	0	PC 17598	31.6833	NaN	S
23	915	1	Williams, Mr. Richard Norris II	male	21.0	0	1	PC 17597	61.3792	NaN	C
24	916	1	Ryerson, Mrs. Arthur Larned (Emily Maria Borie)	female	48.0	1	3	PC 17608	262.3750	B57 B59 B63 B66	C
25	917	3	Robins, Mr. Alexander A	male	50.0	1	0	A/5. 3337	14.5000	NaN	S
26	918	1	Ostby, Miss. Helene Ragnhild	female	22.0	0	1	113509	61.9792	B36	C
27	919	3	Daher, Mr. Shedid	male	22.5	0	0	2698	7.2250	NaN	C
28	920	1	Brady, Mr. John Bertram	male	41.0	0	0	113054	30.5000	A21	S
29	921	3	Samaan, Mr. Elias	male	NaN	2	0	2662	21.6792	NaN	C
30	922	2	Louch, Mr. Charles Alexander	male	50.0	1	0	SC/AH 3085	26.0000	NaN	S



https://www.kaggle.com/competitions/titanic/data

My workflow to tackle the problem





Handling the Missing Data

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

1

Checking the number of missing values

By using isnull() function, I can see that there were missing values for 'Age', 'Cabin' & 'Embarked'

2

Age & Embarked

Replacing the missing values with the mean age of all the passengers for Age.

Since there were only 2 missing values for Embarked, I replaced it with the mode for it.

3

Cabin

As there were too many missing data for Cabin, the decision was made to drop it entirely.

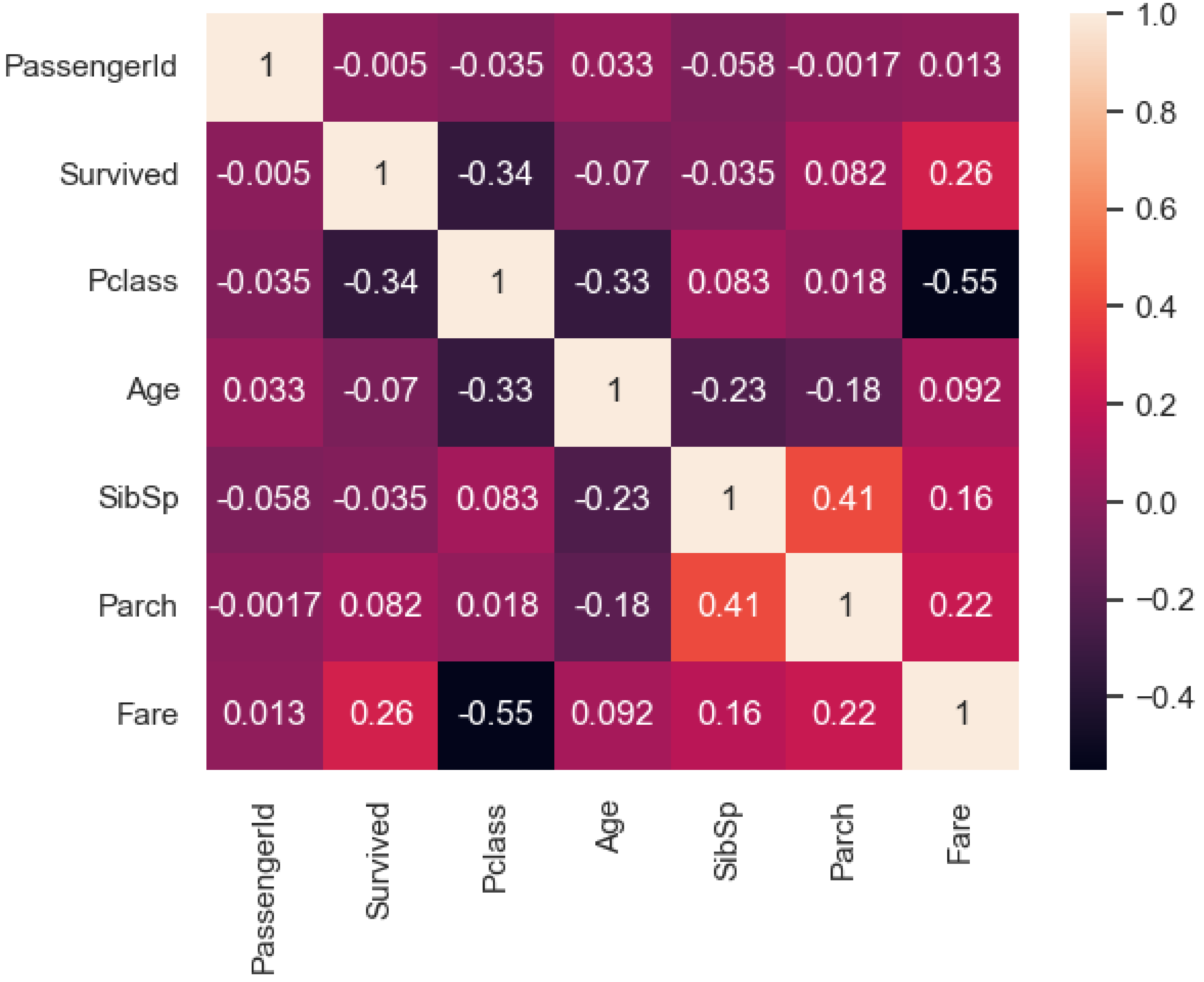
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

4

Checking the number of missing values again

Analysis - Heatmap

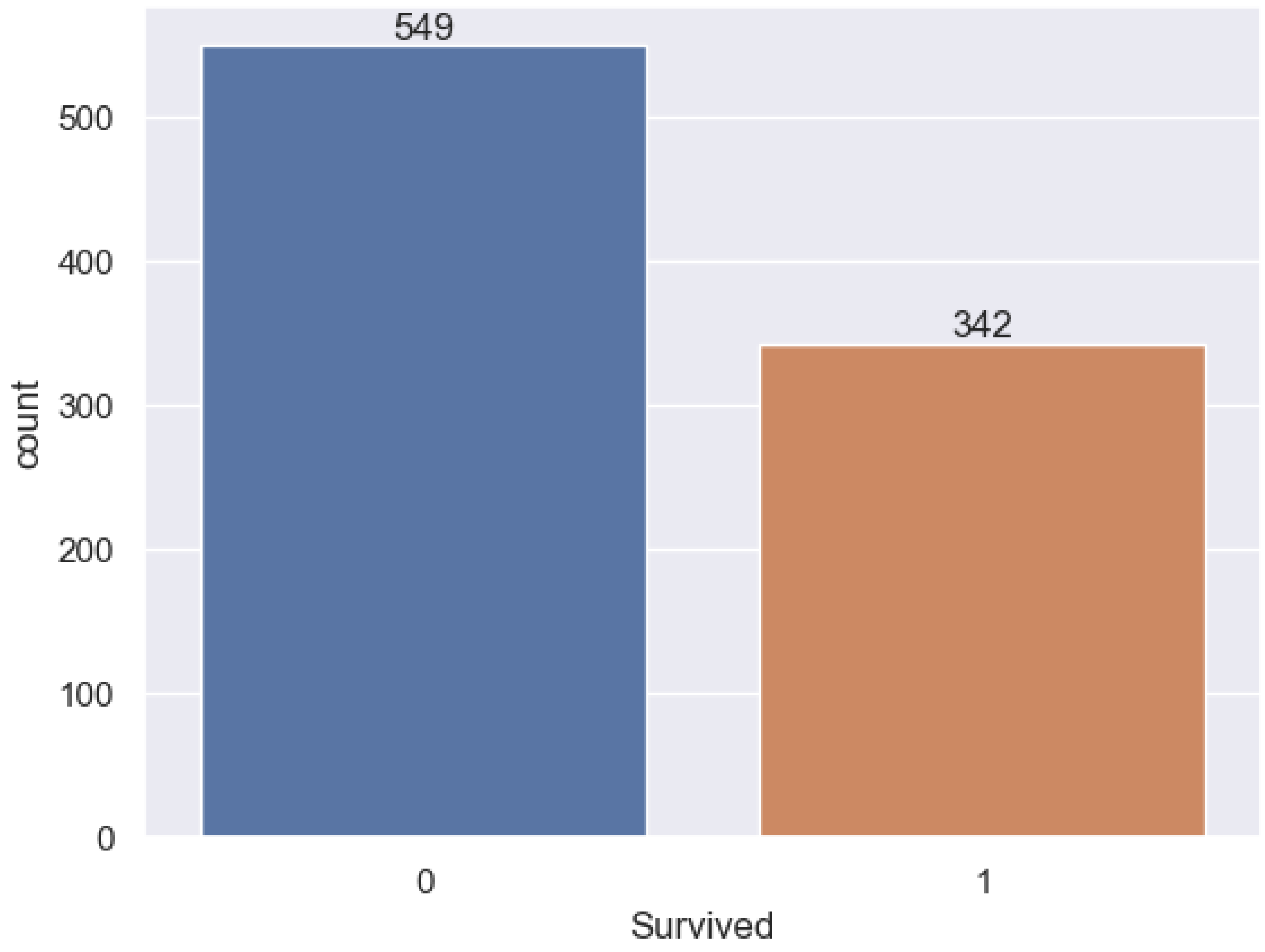
- There is not a good correlation between survived, Pclass and Age.
- There is not a good correlation between Pclass and Survived, Age, Fare.
- There is not a good correlation between Sibsp and Age.
- There is not a good correlation between Parch and Age.

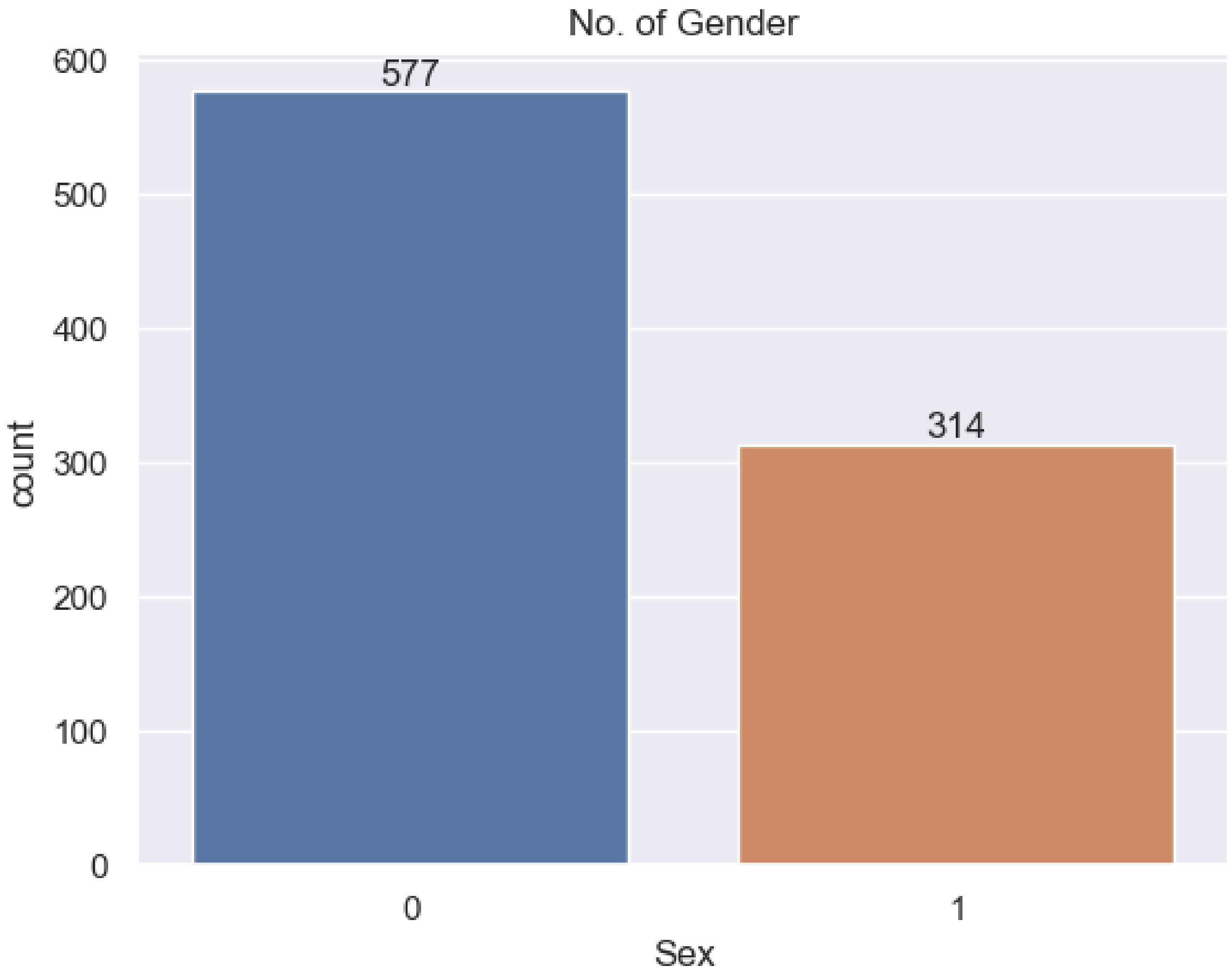


Analysis - Gender

There are a total of 891 passengers.
The number of survivors is 549.
The number of non-survivors is 342

There are 577 Males & 314 Females.

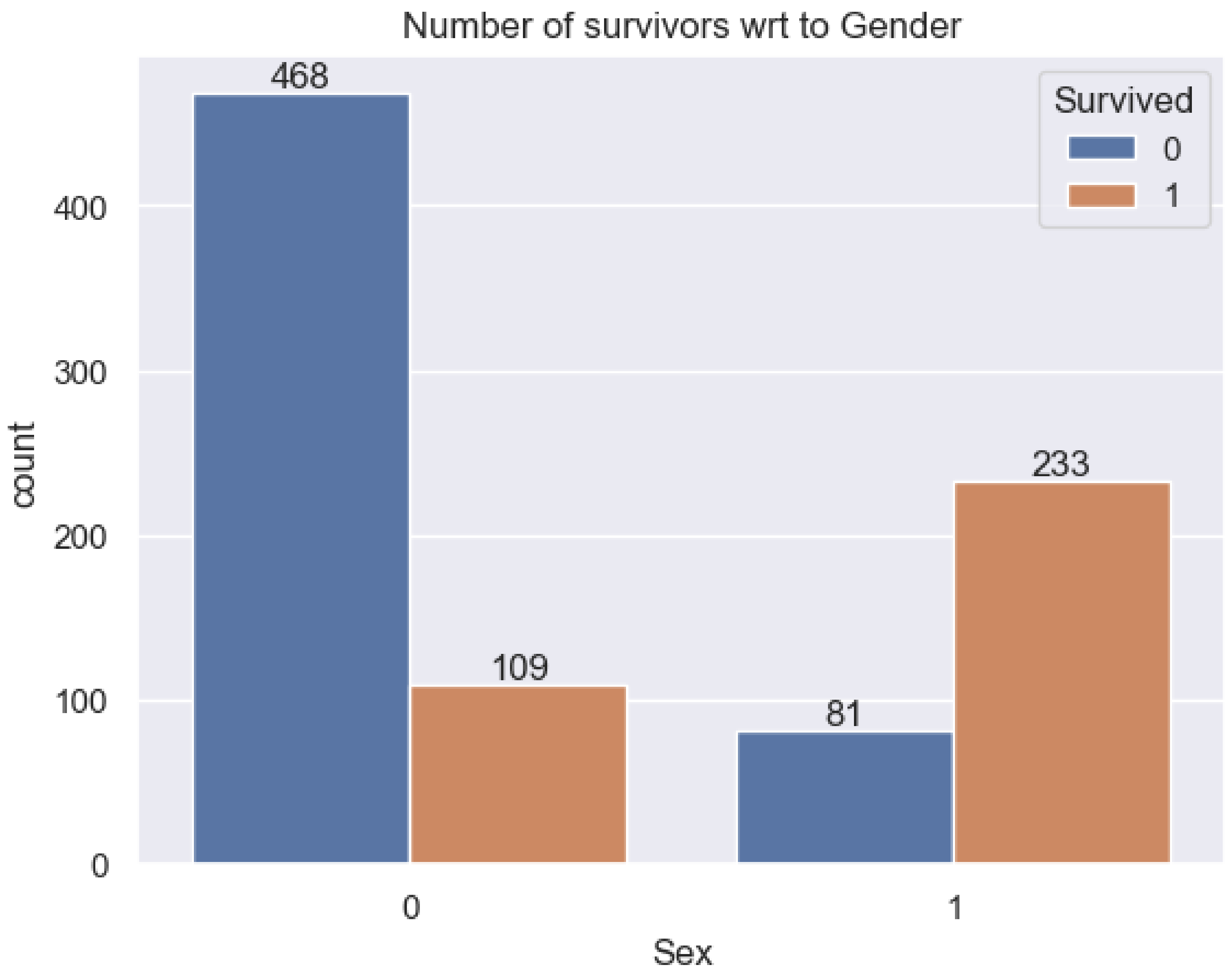




Analysis - Gender

Out of 577 Males, 468 of them did not survive and 109 survived.

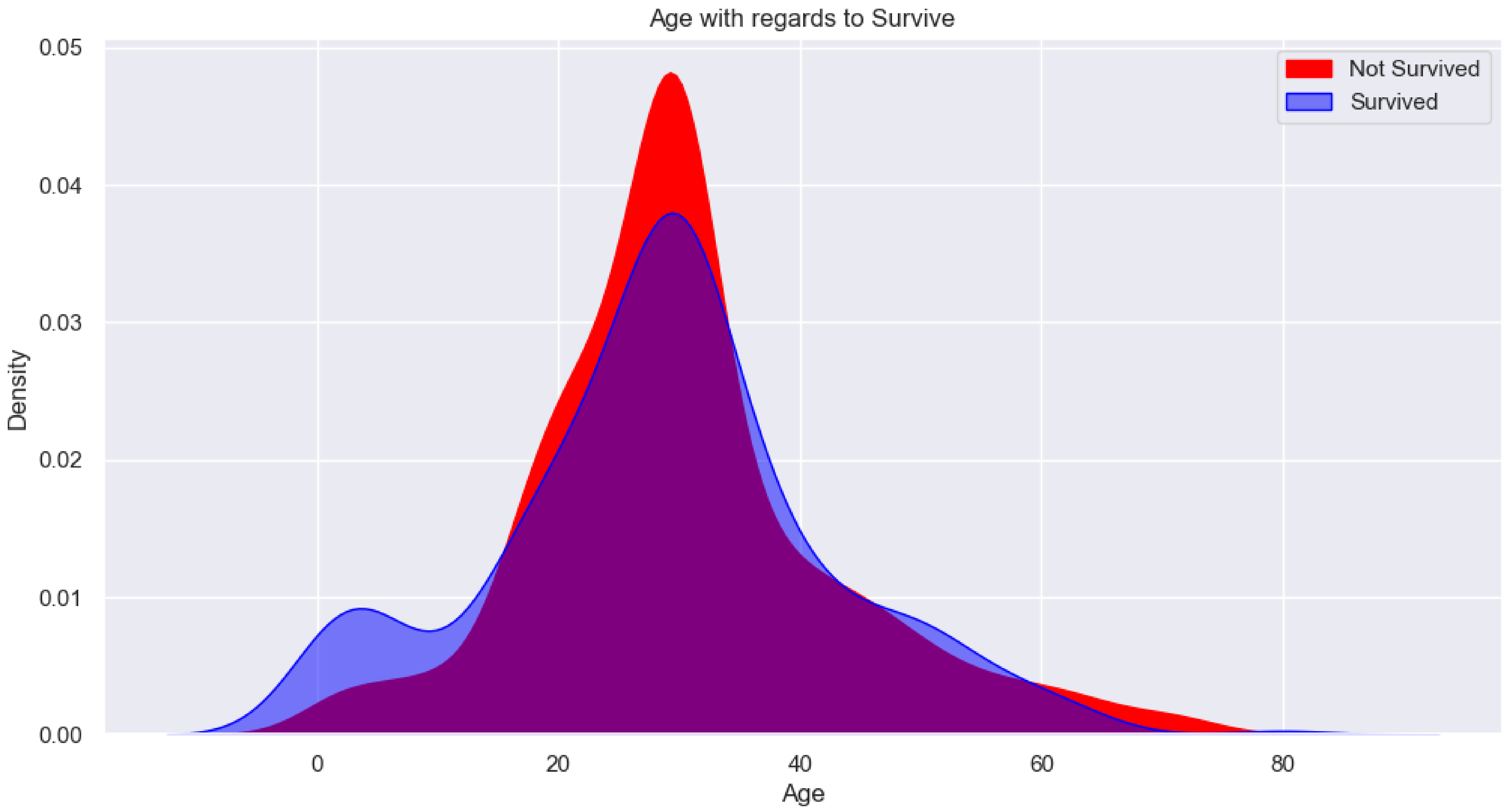
Out of 314 Females, 81 of them did not survive. Whereas, 233 survived.



Analysis - Age

According to the KDE (Kernel Density Estimate):

- Most of the passengers who survived were children & teenagers
- Most of the passengers who did not survived were middle aged adults

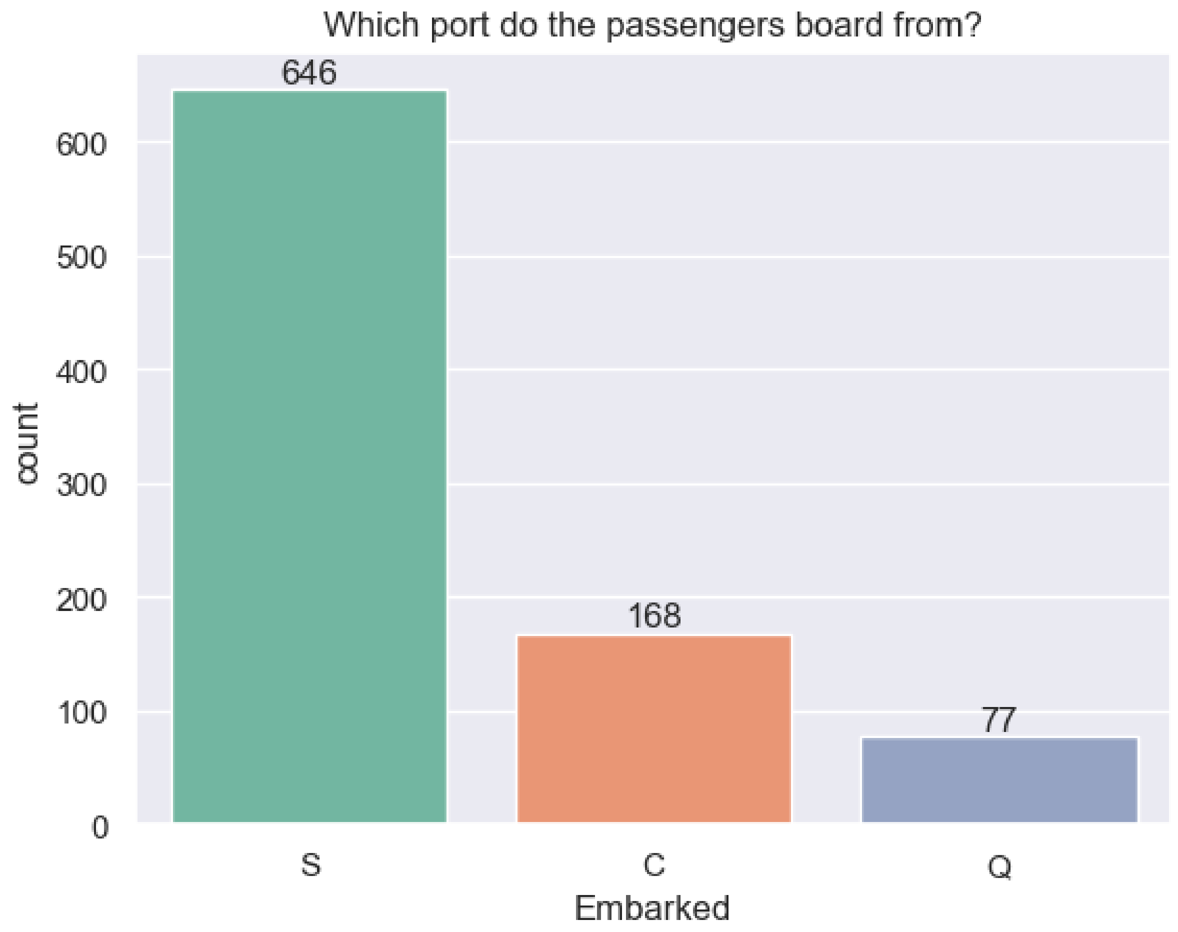


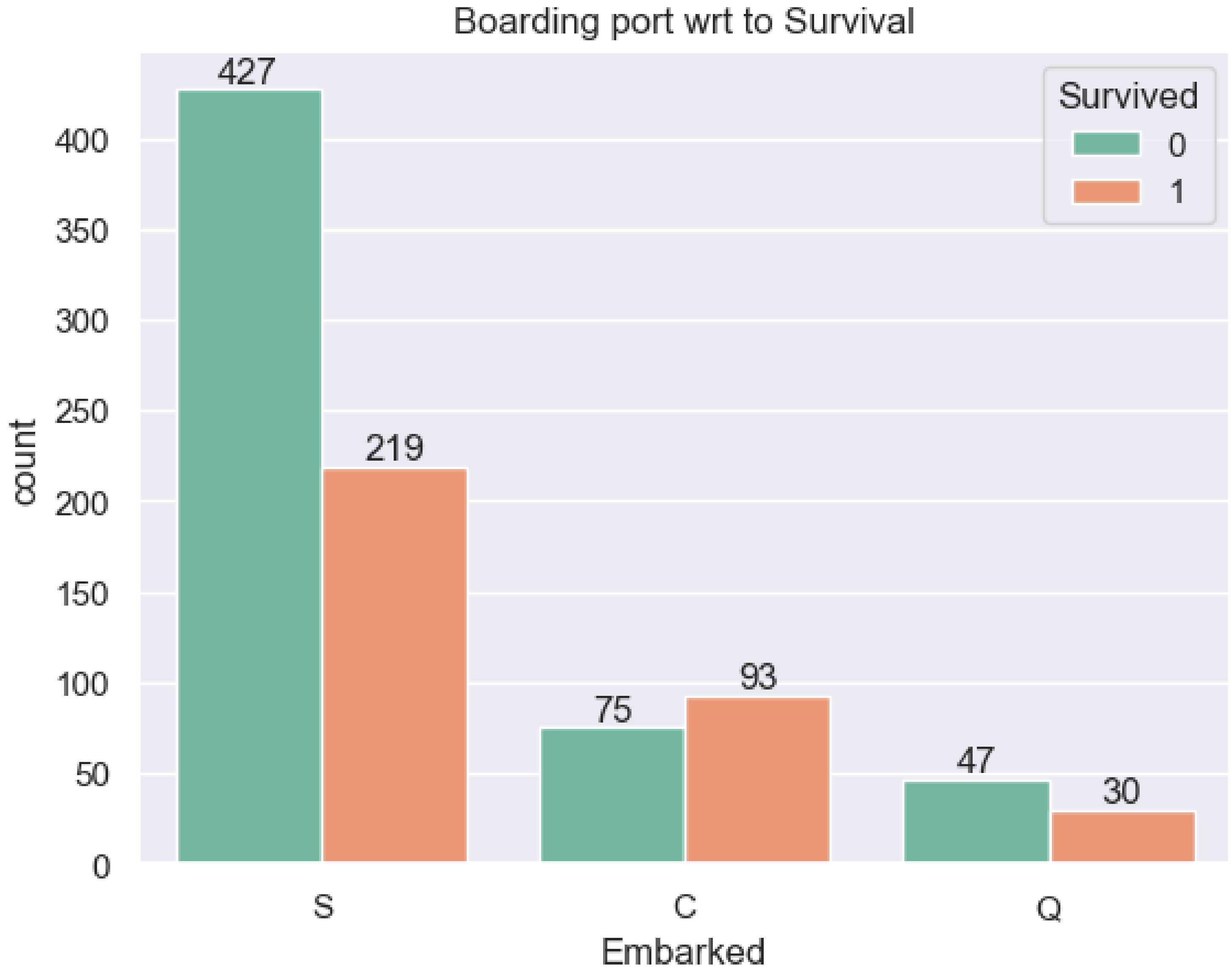
Analysis - Embarked

The three ports were Queenstown, Ireland (present day Cobh), Southampton, U.K, and Cherbourg, France.

646 boarded from Southampton, 168 from Cherbourg and 77 from Queenstown.

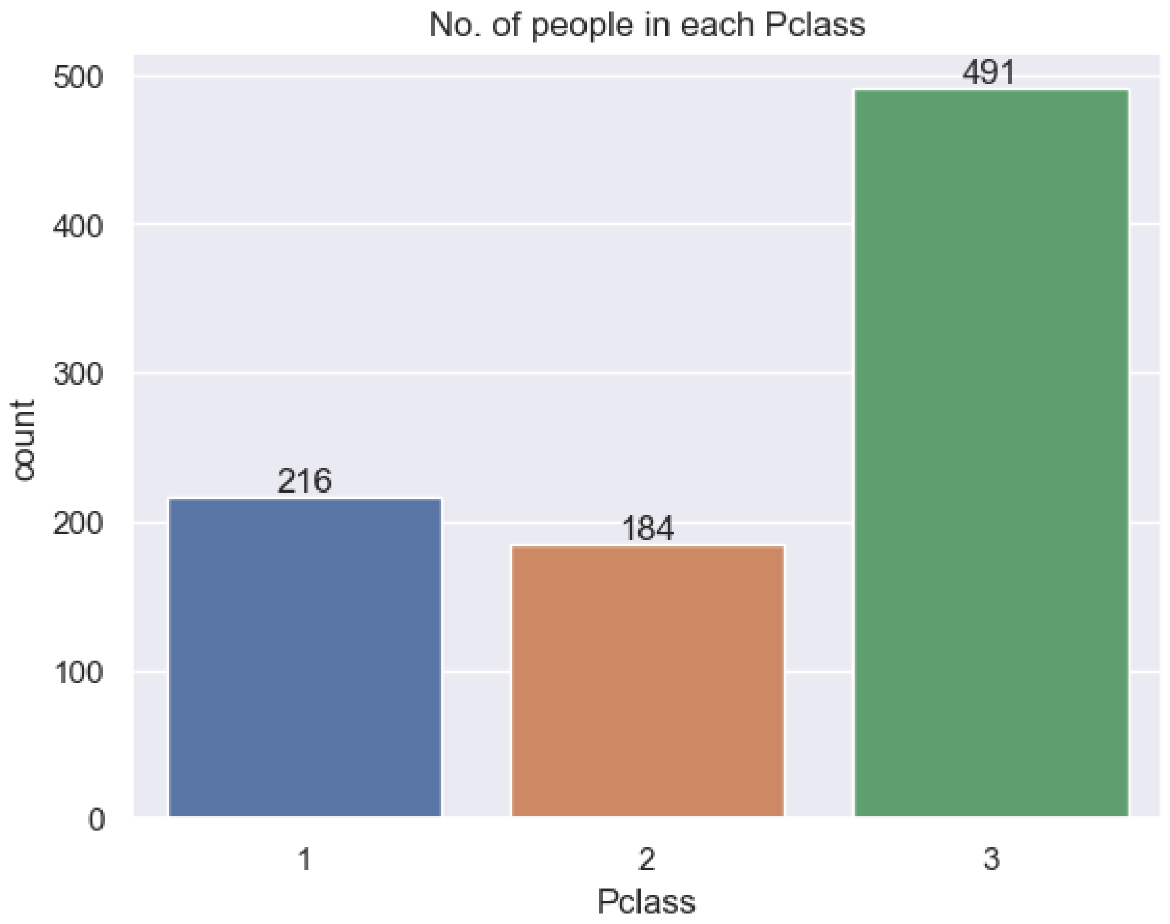
Based on the chart, the highest rate of survival as shown – was from Cherbourg, France where over half of the passengers departing from this region survived the accident.





Analysis - Pclass

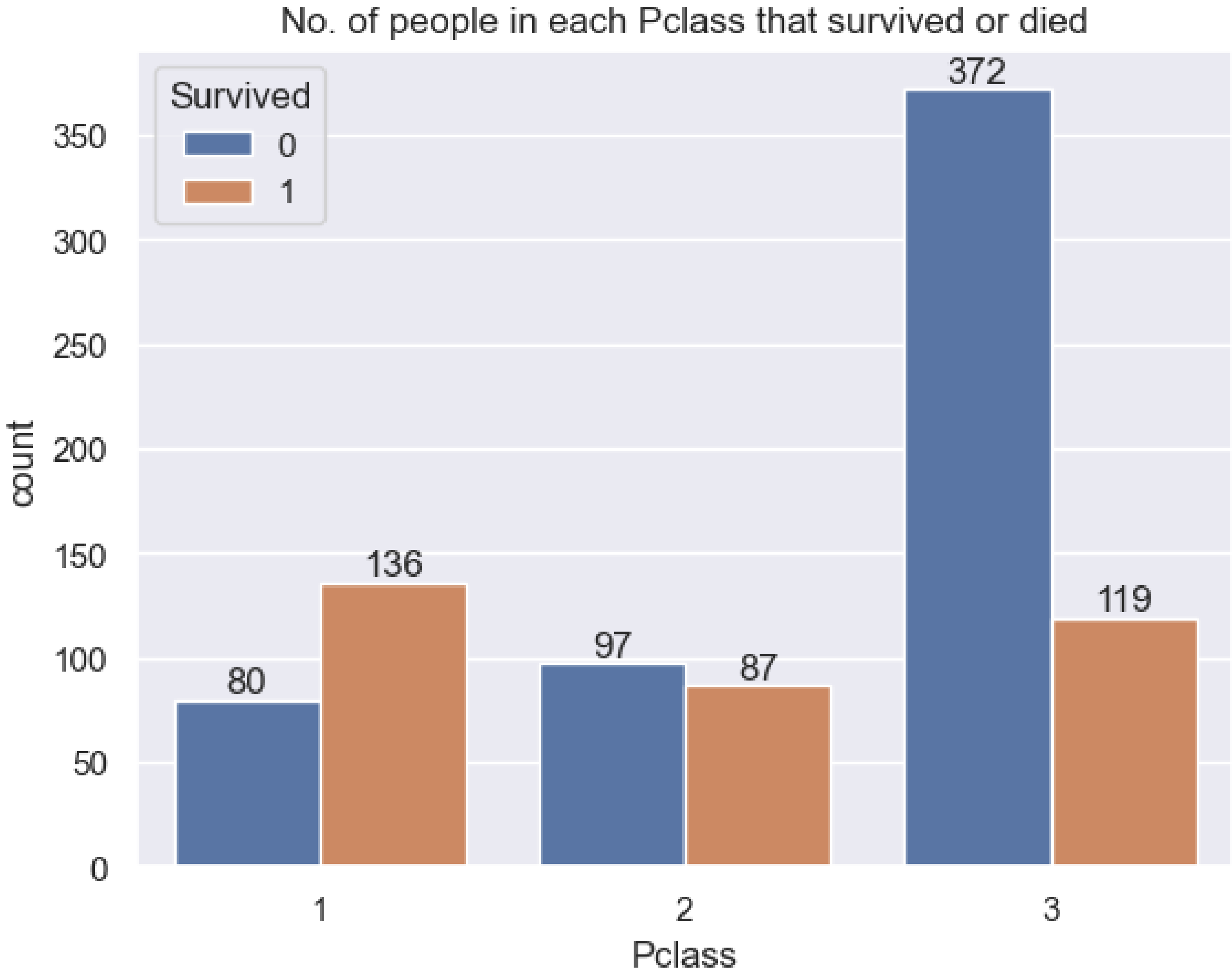
Split by 1st Class, 2nd Class & 3rd Class.
There were 216 passengers in 1st Class, 184 passengers in 2nd Class and 491 passengers in 3rd Class.



The majority of causalities came from the 3rd class, where 372 passengers did not survive.

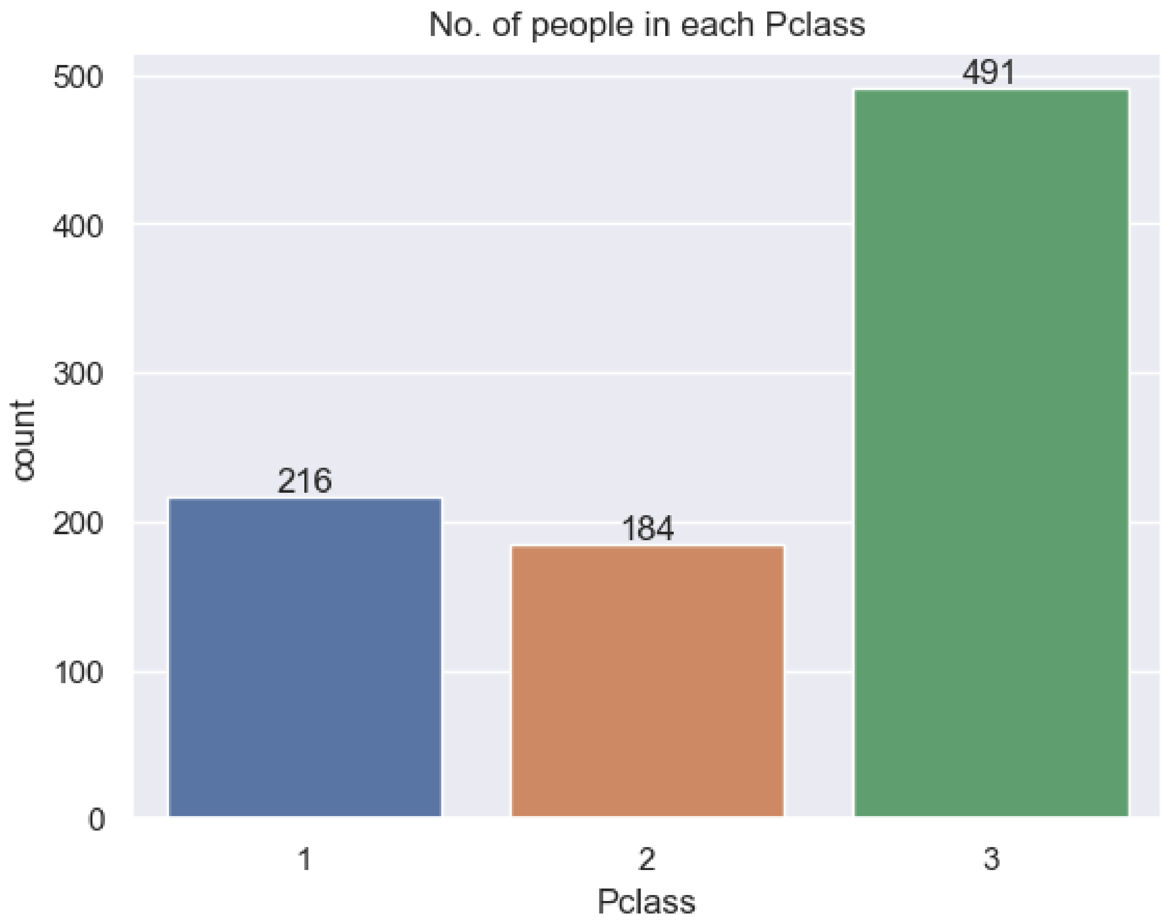
The highest number of survivor came from the 1st class.

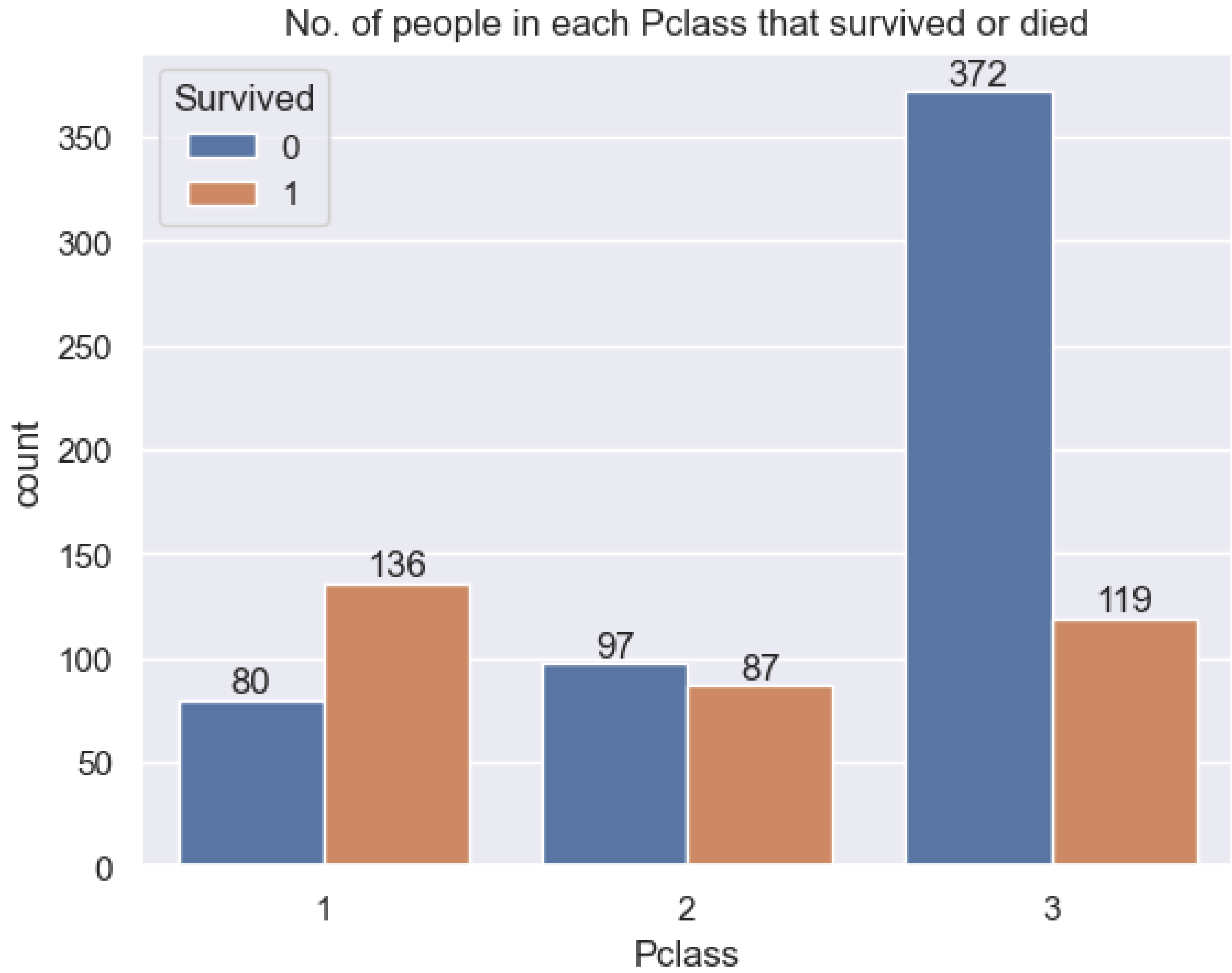
If you were to buy a ticket at the 3rd class, you would have a 76% chance of not surviving the shipwreck.



Analysis - Pclass

```
Pclass = 1
% died = 37.037
The average age who can't survived = 28.78
The average age who have survived = 34.78
The average fare who can't survived = 20.78
The average fare who have survived = 95.61

----------------------------------------------

Pclass = 2
% died = 52.717
The average age who can't survived = 30.09
The average age who have survived = 26.08
The average fare who can't survived = 33.3
The average fare who have survived = 22.06

----------------------------------------------

Pclass = 3
% died = 75.764
The average age who can't survived = 30.7
The average age who have survived = 23.23
The average fare who can't survived = 35.06
The average fare who have survived = 13.69
```

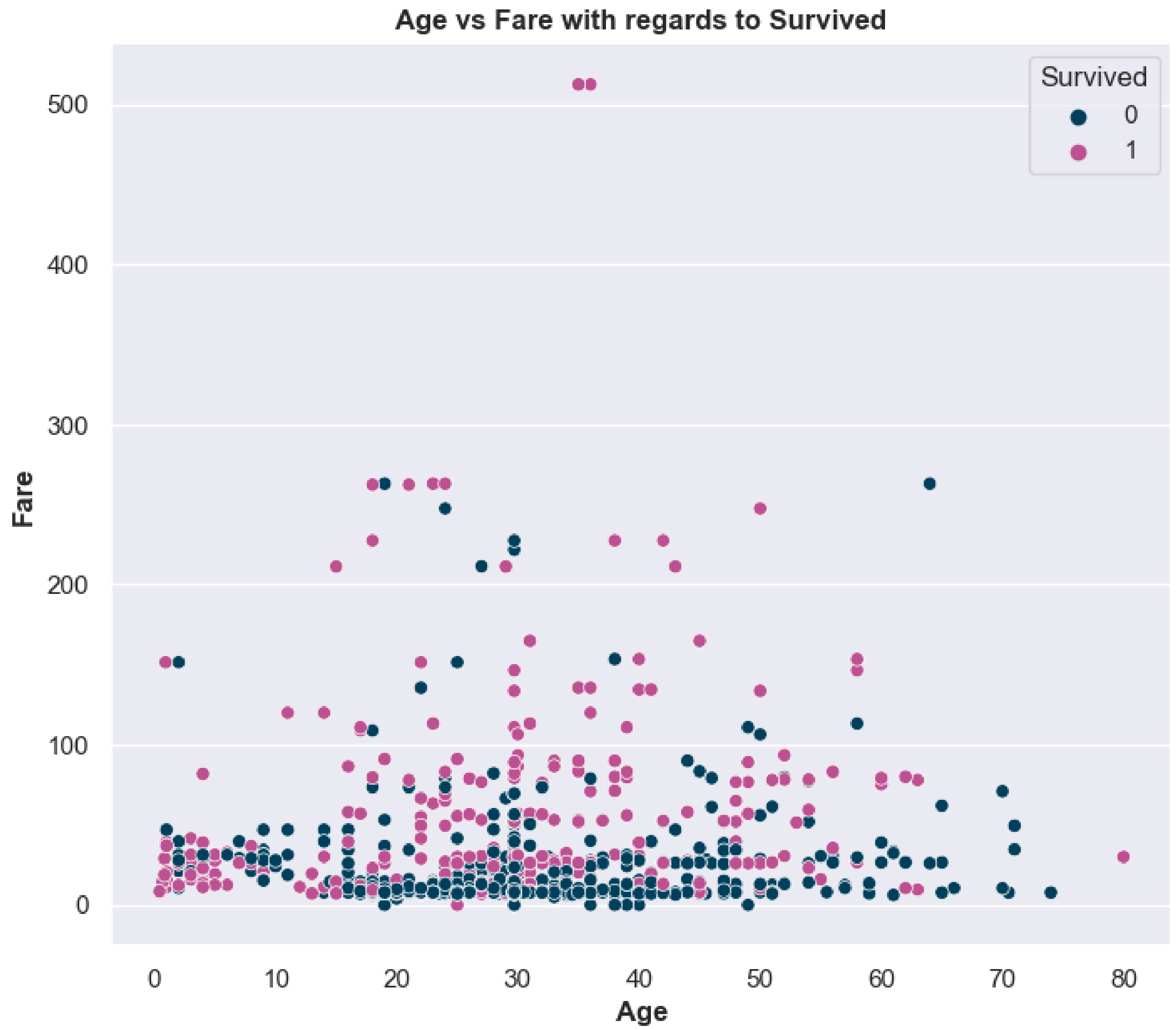


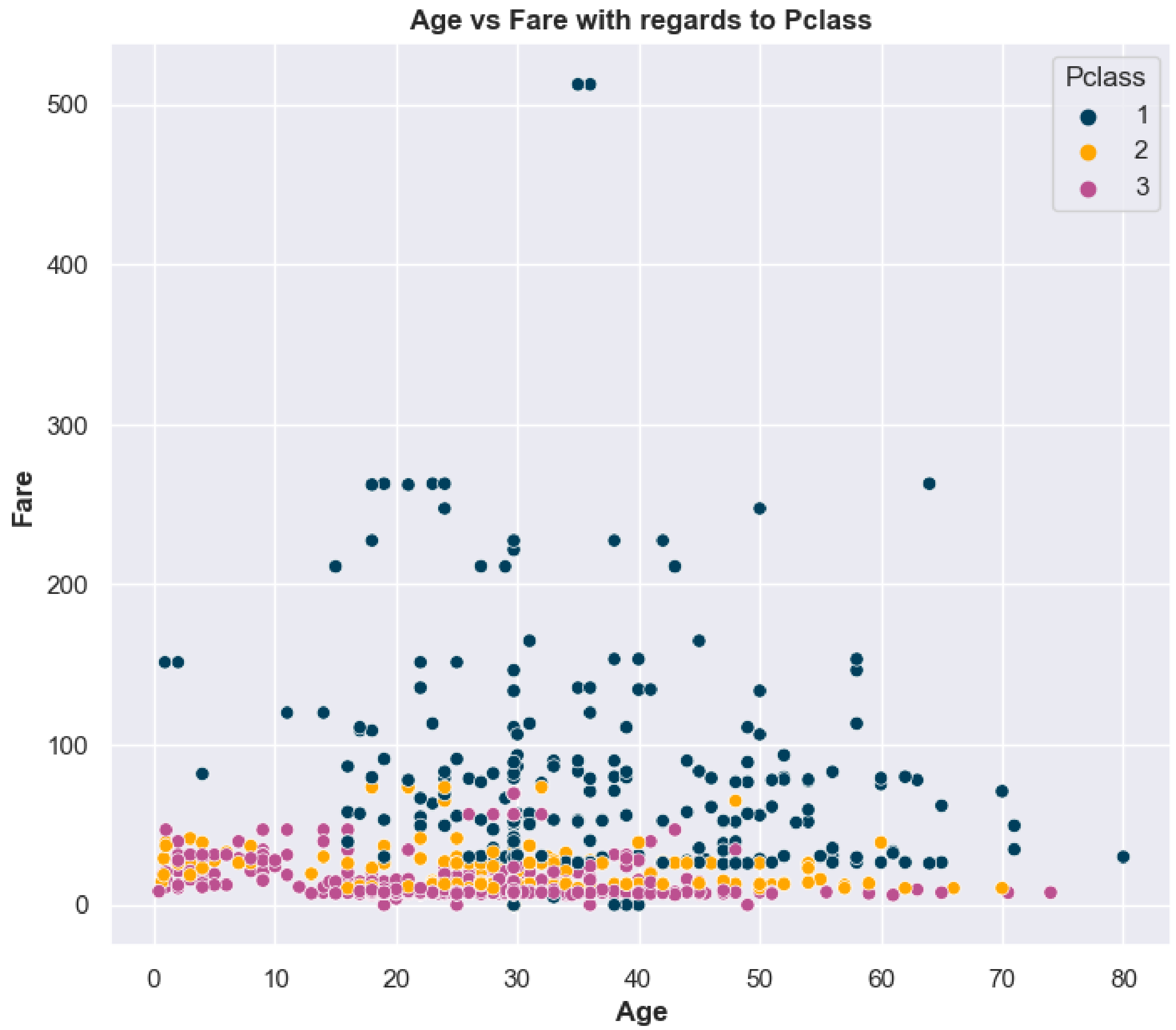


Analysis - Age vs Fare

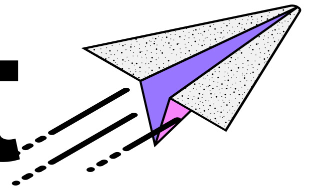
Children and Teenagers had survived. Fare could not have possibly affected these group.
However, it shows that as age and fare increases, there is a difference with passengers who had low fare ticket as they did not survive and vice versa.

Passengers with high fare tickets who are mostly from 1st class had survived.





Before doing train test split



1

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S

Encoding the categorical data

The current data set still contains categorial data. As most machine learning models only works with integer values, I have converted the affected columns to integer format.

2

```
#converting the categorical columns
df_titanic.replace({'Sex':{'male':0,'female':1}, 'Embarked':{'S':0,'C':1,'Q':2}}, inplace=True)
✓ 0.3s
```

Checking encoded columns

3

```
X = df_titanic.drop(columns = ['PassengerId', 'Name','Ticket','Survived'],axis=1)
Y = df_titanic['Survived']
✓ 0.2s
```

Separating the features and the target

4

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2,random_state=123)
✓ 0.2s

print(X.shape,X_train.shape, X_test.shape)
✓ 0.2s

(891, 7) (712, 7) (179, 7)
```

Split the data into training & testing data

Model #1- Logistic Regression

By doing the accuracy score for the training & testing data, we can see that the accuracy score is similar to each other.
79% & 81% respectively.
This shows that overfitting did not occur.

The AUC for the ROC graph is 0.81, it is considered to have excellent discrimination and performing well.

```
[1 0 0 0 1 0 1 0 1 0 1 0 0 1 0 1 1 0 0 1 0 1 1 0 0 0 0 0 1 1 0 0 1 1 0 0 0
 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0
 0 0 1 0 1 1 1 0 1 0 0 1 1 0 1 1 1 0 0 0 1 0 1 1 1 0 0 1 0 0 1 0 1 0 0 0 1
 0 1 0 1 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0
 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 1 1 0 1 0
 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 1
 0 1 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 1 0 0 1 1 0 1 0 1 0 0 1 1 0 0 1 0 0
 1 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
 0 0 1 1 1 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 1 1 1
 0 0 1 1 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0 1 1 0 1
 0 0 0 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 1 1 0 1 0 1 1 1 0 0 1 0 1 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 0
 0 0 0 0 0 1 0 1 1 0 1 1 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0
 0 0 1 1 1 1 1 1 1 0 1 0 0 1 1 0 0 0 1 1 1 1 1 1 0 0 1 0 0 0 0 0 0 1 0 0 0
 0 1 0 0 1 0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1
 0 0 0 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1 1 0 1 0 0 1 0 1 0 0 0 0 1 1 0 0 1 0 0
 0 0 0 1 0 0 1 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 1 1 1
 0 1 0 1 1 0 0 1 0 1 1 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 1 0 1 1 0 1 0 0
 0 1 0 1 0 0 0 1 0 1 1 1 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 1 0 1 0 0 1
 1 0 0 1 1 1 0 0 0]
```

training data prediction

```
The accuracy score of the train data is: 0.7949438202247191
```

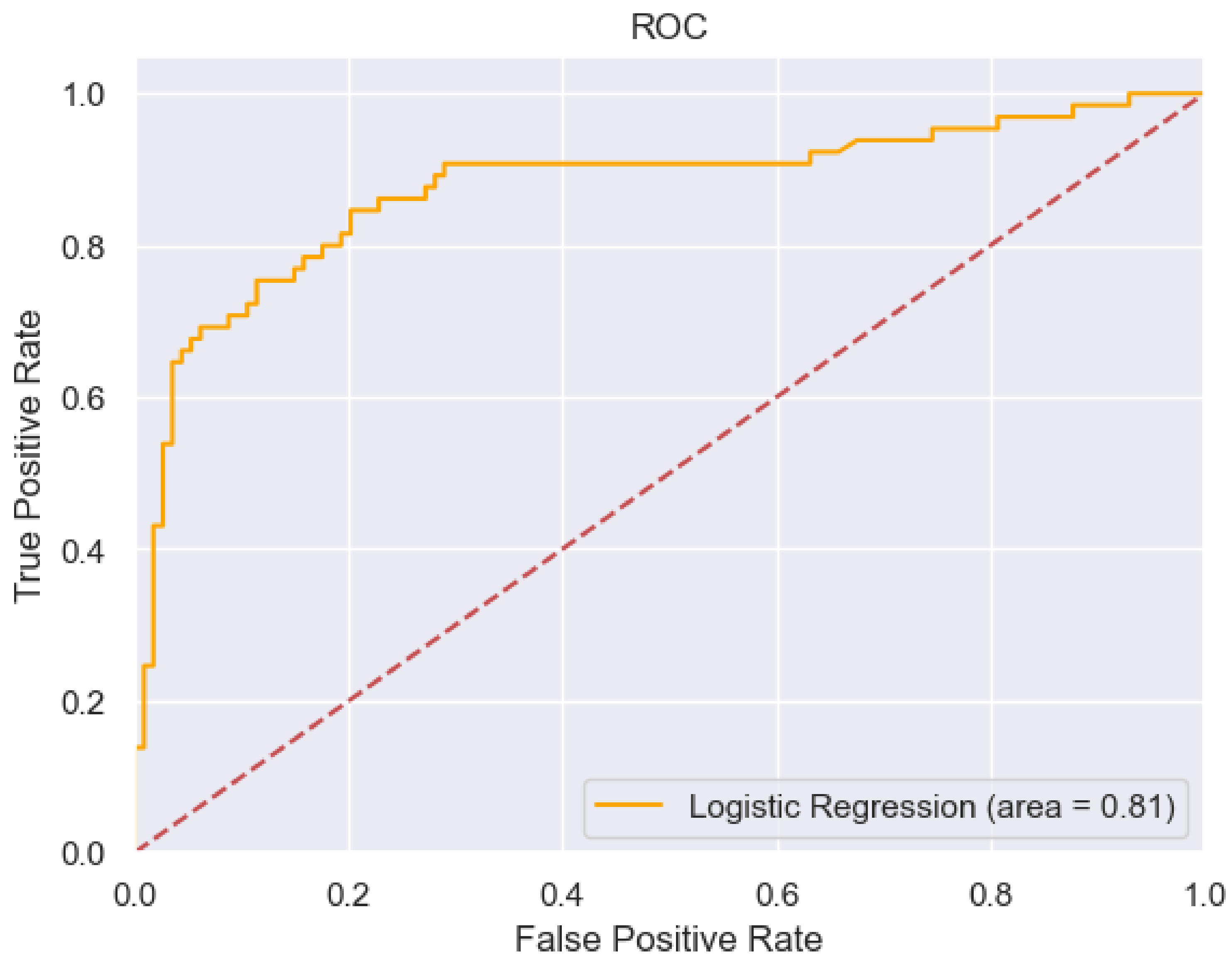
Accuracy Score: 79%

```
[1 0 1 0 0 0 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0
 0 1 0 1 0 1 0 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 0 1 0
 0 0 0 0 1 1 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 1 1
 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 1 0 0 1 1 0 1 1 0 1 0 1 0 0 1 0 0 0
 0 1 0 0 1 0 1 0 1 1 1 1 1 0 1 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 0]
```

test data prediction

```
The accuracy score of the test data is: 0.8156424581005587
```

Accuracy Score: 81%



Model #2 - Decision Tree Classifier

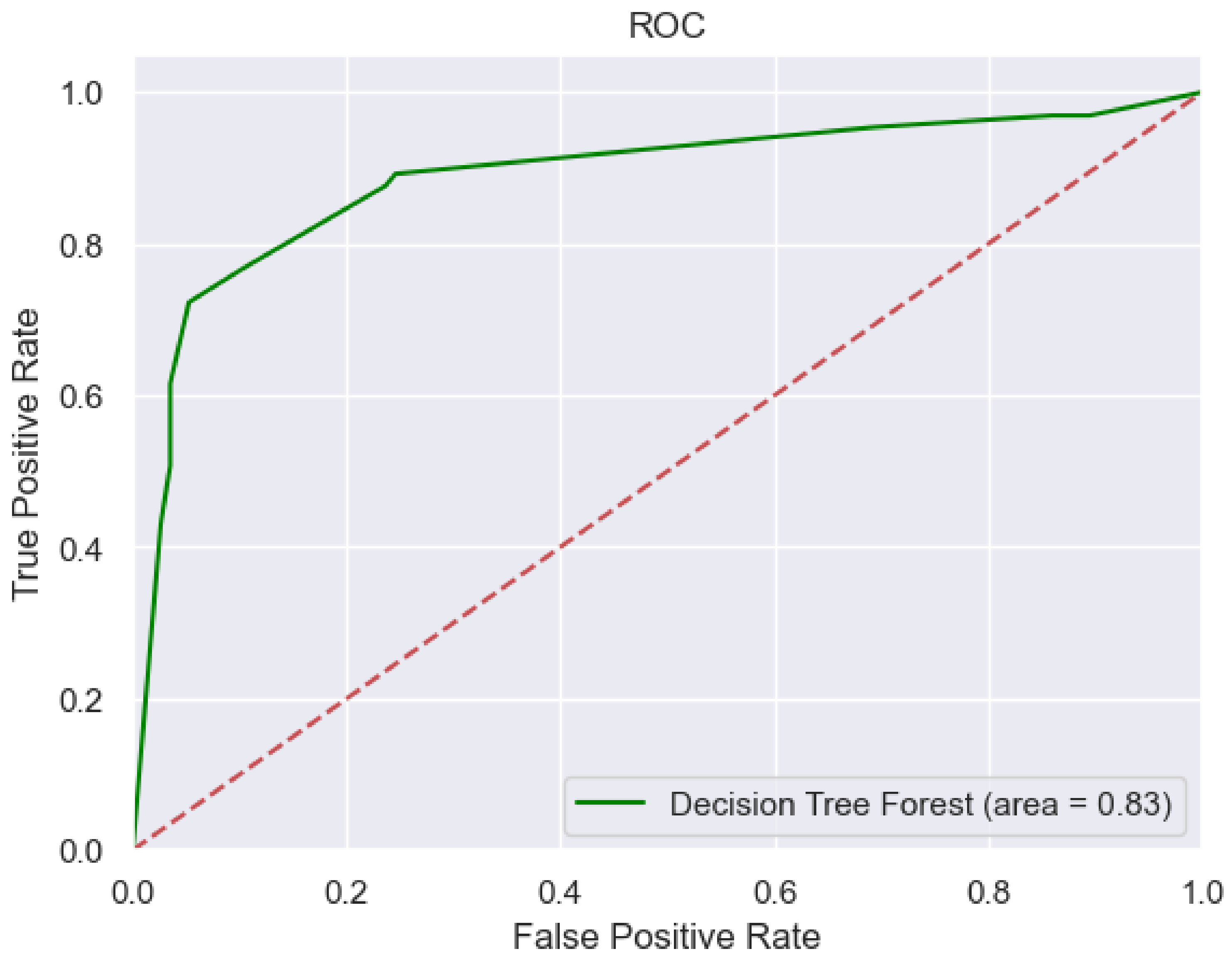
With a max_depth of 5, the accuracy score of the model is 85%

The AUC for the ROC graph is 0.83, it is considered to have excellent discrimination and performing well also.

```
The accuracy score of the test data is: 0.8491620111731844
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       114
           1       0.81      0.77      0.79        65

    accuracy                           0.85       179
   macro avg       0.84      0.83      0.84       179
weighted avg       0.85      0.85      0.85       179
```

test data prediction

Accuracy Score: 85%



Model #3 - Knearest Neighbour

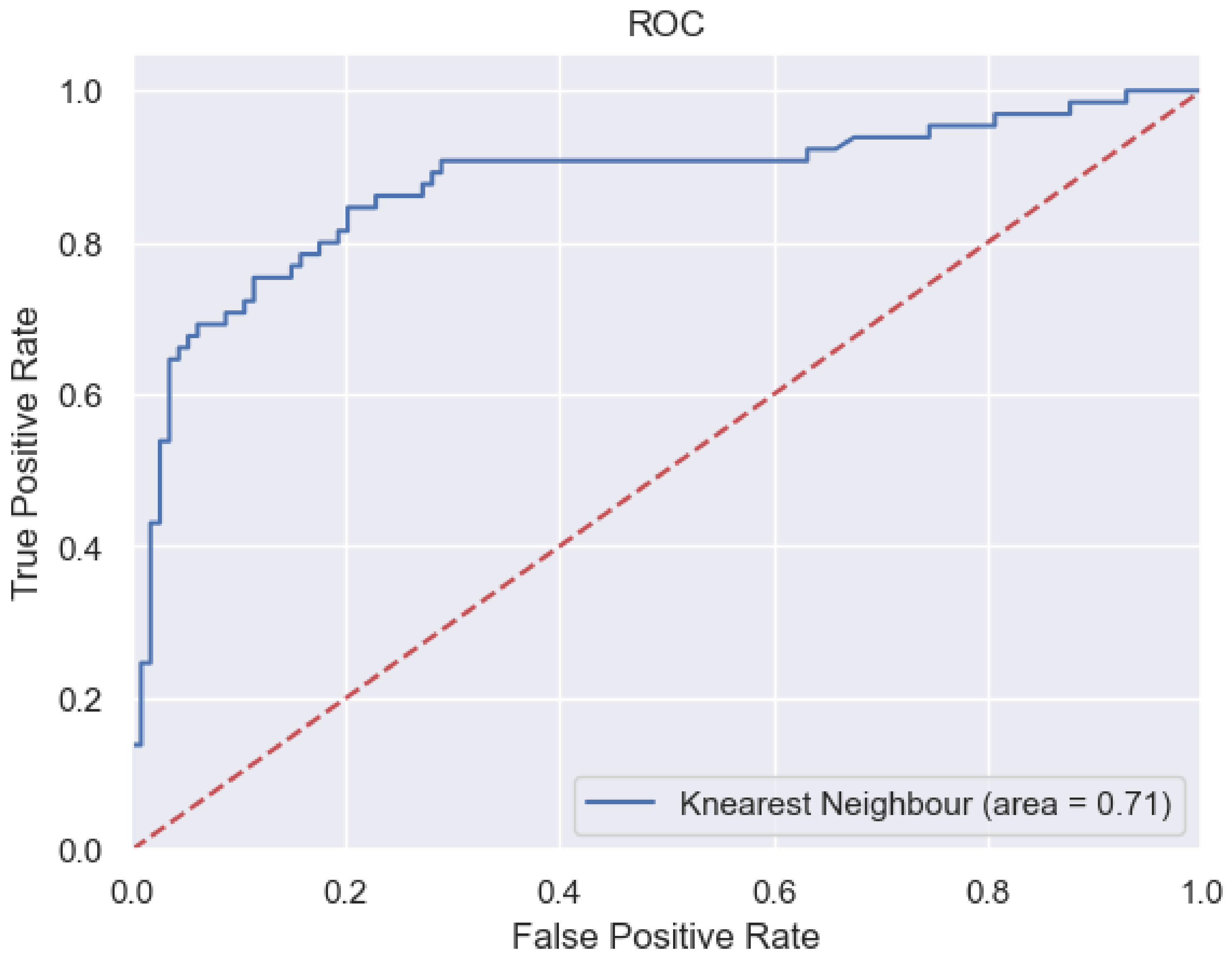
With parameter n_neighbour of 3, the accuracy score of the model is 73%

The AUC for the ROC graph is 0.71, it is considered to have acceptable discrimination and performing fairly.

```
The accuracy score of the test data is: 0.7318435754189944
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       114
           1       0.63      0.63      0.63        65

    accuracy                           0.73       179
   macro avg       0.71      0.71      0.71       179
weighted avg       0.73      0.73      0.73       179
```

test data prediction

Accuracy Score: 73%

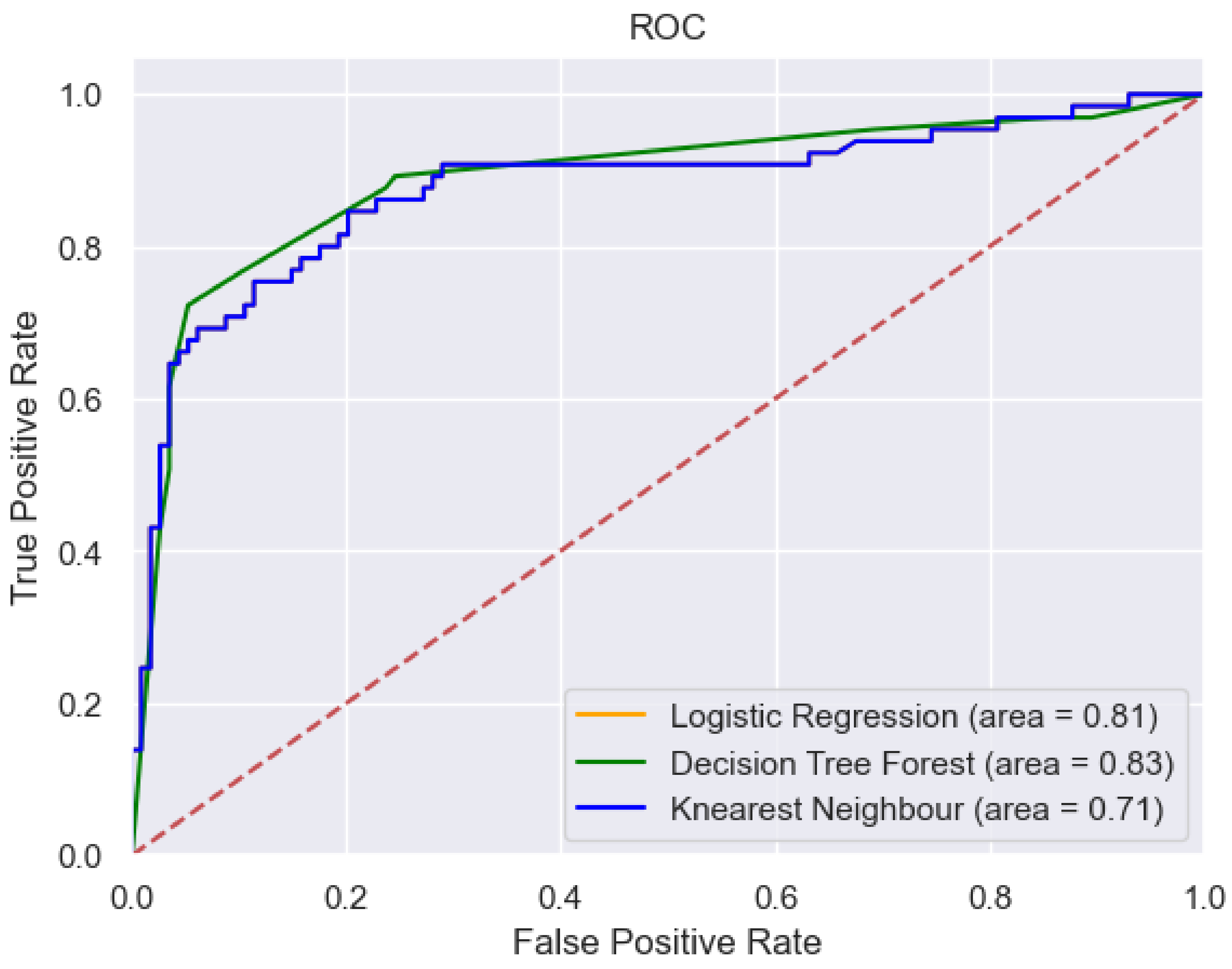


Evaluation

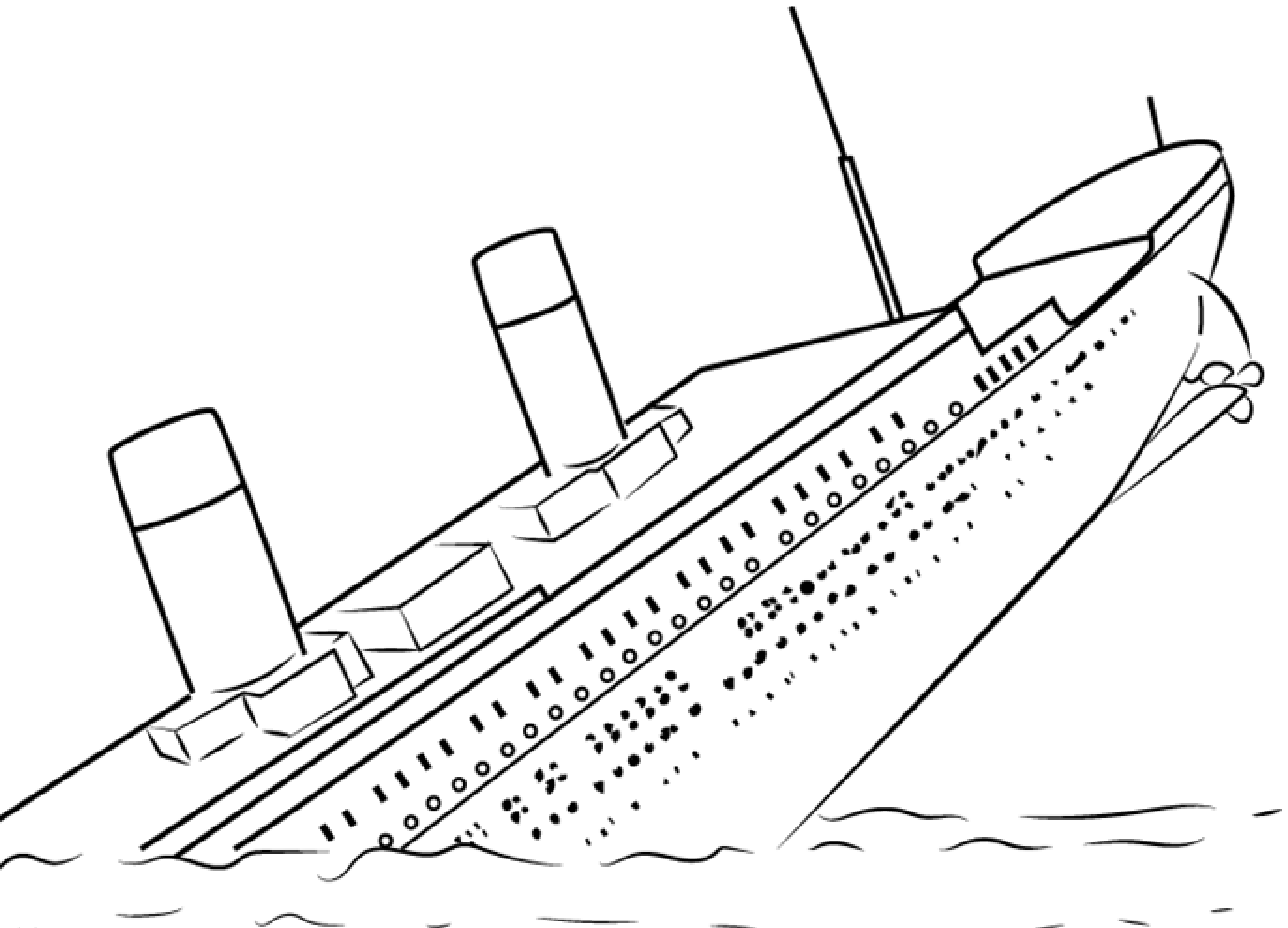
Passengers who had higher chance of survival were:

- Children & Teenagers
- Embarked from Cherbourg
- Had high fare tickets

By comparing the 3 models, KNN has the lowest accuracy score and AUC. Though Logistic Regression had excellent accuracy score and AUC. However, decision tree model has the best accuracy score as well as AUC score. Thus, decision tree model is the best in predicting the survivability of the passengers of titanic.



Thank you!



https://github.com/spreadjam/Machine-Learning-Titanic